**University of Canterbury**

*DATA 422 – 20S2*

Group Project

Lecturers: Giulio Dalla Riva and Thomas Li

Names and Student Ids:

Introduction (from 1 to 7.1) – Thong Minh Nguyen (Martin) – 68623131

Andy Huynh – 64499983

Sakthi Harish Thennarasu – 64916650

Nur Syaida Firzana Mohamad Isham – 66474666

MARTIN NGUYEN – GDELT WEBSITE WEB-SCRAPING WITH R AND JULIA

1. **Brief description of the project**

The purpose of the project is to observe how the mass shootings were represented in the articles in America based on tone and volume intensity through the period of time from 2018 to 2020. There were too reasons for choosing this topic. One is that the mass shootings have been a serious problem in the United States that one of them was ranked as the deadliest shooting event and they have been ocurring approximately every 12.5 days. The other is that for this major problem, how the articles mentioned about mass shootings is also the part that we want to discover, as this would affect the readers in terms of ethics. Hence, we would like to provide the data about the mass shootings infromation (number of people died) and how they were represented in the articles (the tone and the volume intensity). After that, the data would be presented under the two sections as following:

* The tables includes the date that the event occurred, the average tone values, the volume intensity values, URLs of the top articles, and the names of the articles.
* The charts visualises the average tone, the volume intensity values and total number of people died and injured through the date, and the tone chart.

The first one would include two main parts of the project – webscraping and wrangling data. The result from this could be provided publicly for people who might be in need for the crime text analysis. The second part would be seen as some examples for what information the final tables could provide.

1. **The scope and some terms**

Our project would focus on the mass shootings and articles within USA from 01/01/2018 to 09/10/2020 (3 years). The definitions of some term used in the project were listed as below

* Average tone: We would focus on the average tone of the articles of the day after the incident
* Volume intensity: We would focus on the percentage between the number of the mass shooting articles and the total articles of the day after the incident
* Tone histogram: We would focus on the distribution of tone of mass shooting articles for each year.

1. **What data sources you used**

The data were collected based on two main parts

* Part 1: The data about the articles including date, tone, volume intensity and information of articles (url and names) relating to mass shootings.
* Part 2: The data about the date and information about that mass shooting s(location, number of people get injured and died, and description)

As to part 1, the data was collected from GDELT project website or Global database of society. This website monitors the broadcast, print, and web news from almost around the world and provides the information about people, locations, organizations, themes, and so on.

As to part 2, the data was collected from wikipedia about the list of the mass shooting from 2018 to 2020.

From this, we combined together to have a final dataset that mainly represents the time, the location, the number of victims, the tone values, the volume intensity values, the description, the urls and the titles. Additionally, the tone histogram data would be set separately from the main dataset, but it would be seen as one of our final output. Through this, we would know about how the media (mainly newspaper) mentioned about mass shooting through times around America based on tone and the number of articles.

1. **Why you chose those data sources**

As to the Gdelt project, there are three main reason for choosing this websites. The first one is that this is the reliable website that collecting the article informaiton from around the world based on the Google Jigsaw. Secondly, the scope of information that this website can cover is considerably huge and informative nearly every corner of every country in over 100 languages. Thirdly, this is the realtime network diagram and database of global human society for open search that contaning the data up-to-date. Hence, based on these reasons, the data we collected from this Gdelt Project would be close to accuracy, much informative, and up-to-date.

As to the information about the mass shootings in the United States, there are three well-known websites that we can collect. They are Gun Violence Archive, Mother Jones and Wikipedia. The reason that wikipedia was chosen is presented as below:

* The Gun Violence Archive provides all the information about the mass shootings including date, number of people died or injured and location. However, the description about the mass shooting was not mentioned, and this information would be neccessary for people who would use our project dataset.
* The Mother Jones provides all the information as The Gun Violence Archive and also the descriptions in details. However, this website was not recorded enough number of shootings each year like The Gun Violence Archive.
* The information from wikipedia is the combined between the data from Gun Violence Archive and Mother Jones (this was mentioned under refernce parts so it is reliable). Hence, this includes nearly all the dataset the we needed for the mass shootings information.

1. **Targets**

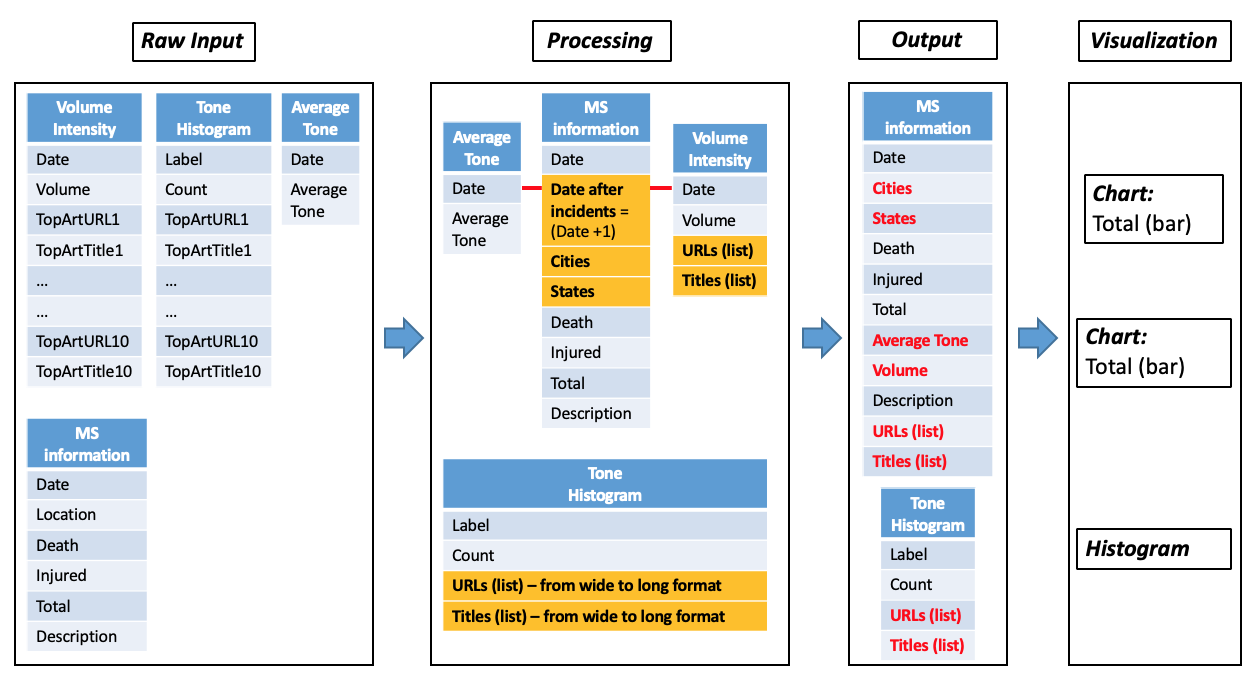
As to the final model of database, for each key words we used for the term mass shooting, there would be two types of final tables as described below

* A table including dates of the incident, the number of people get affected (died, injured and both), the location, the description, the average tone value, the volume intensity value, the top links (urls) of the articles, and the name of the articles (could be separate for each year)
* A table for each year listing the average tone value labels, the number of articles for each label, the top links (urls) of the articles, and the name of the articles.

From those final tables we would visualise them following the years. Based on these charts we could see how the articles talking about mass shootings in America.

As mentioned above, the results from this could be used for criminal text analysis, and we would provide some examples with visualisations.

1. **The process of the project is presented as below**



The raw input would be scraped from Gdelt Project and Wikipedia by Martin and Andy. The processing data part (modifying the dataset) and visualisation would be done by Sakthi and Syaida.

1. **Project details**
   1. **Web scraping from Gdelt project website**
      1. **Challengings**

Every part in the project would have its own difficulties, but in this case of scraping data from Gdelt Project, I would consider difficulty as the term mentioning about the time-consuming tasks. Following that, searching the data about the mass shooting information (like tone, volume intensity, articles, urls, etc..) in Gdelt project website took us more than one week to know where and how we should extract the data.

During the time of searching this data, we have tried functions from the gdeltr2 library and this was the most time-consuming part. In particular, we tried the functions in package, but the data we obtained from this were either too large to be manipulated even a every simple action like copying or viewing the data, or not enough information to be analysed.

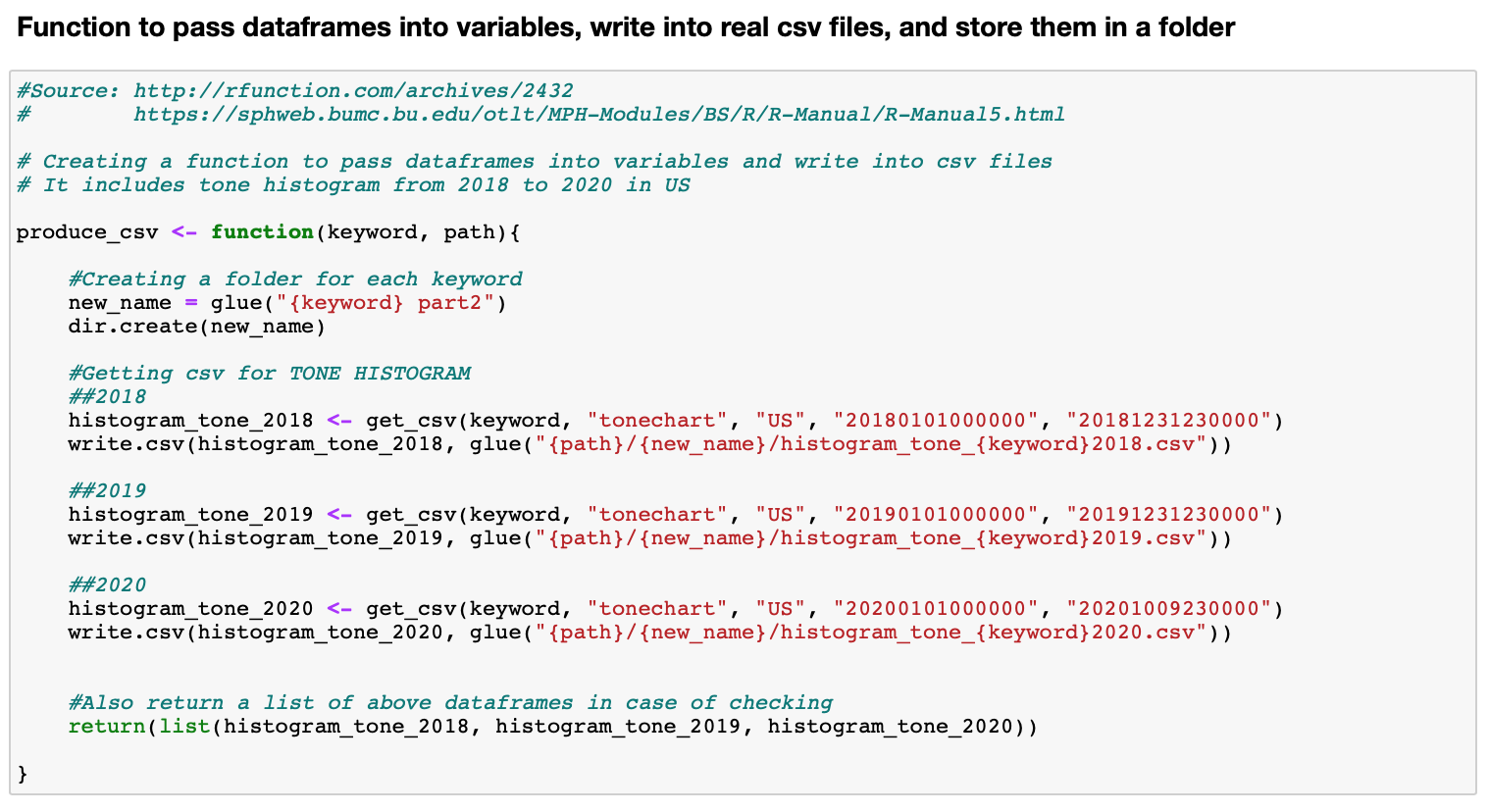
After that time, we found that API was the second way that helped us get the data we need for analysing. To know how to use the Gdelt project API, we spent more than 2 days to understand thoroughly and to know how to develop the script to scrape the data.

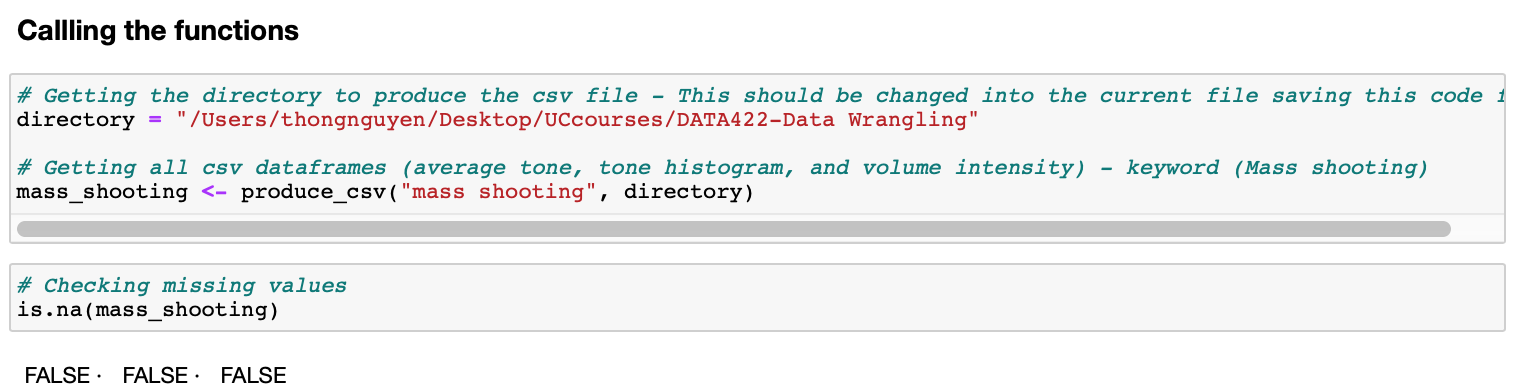
Besides, during the time we were doing the project, Gdelt Project shut down their website for around 2 times and a whole day for each time without any notice. At that time, we did not know what we should do to solve this issues, as we had not saved the files yet and could not extract the data again. However, the website works again after those times.

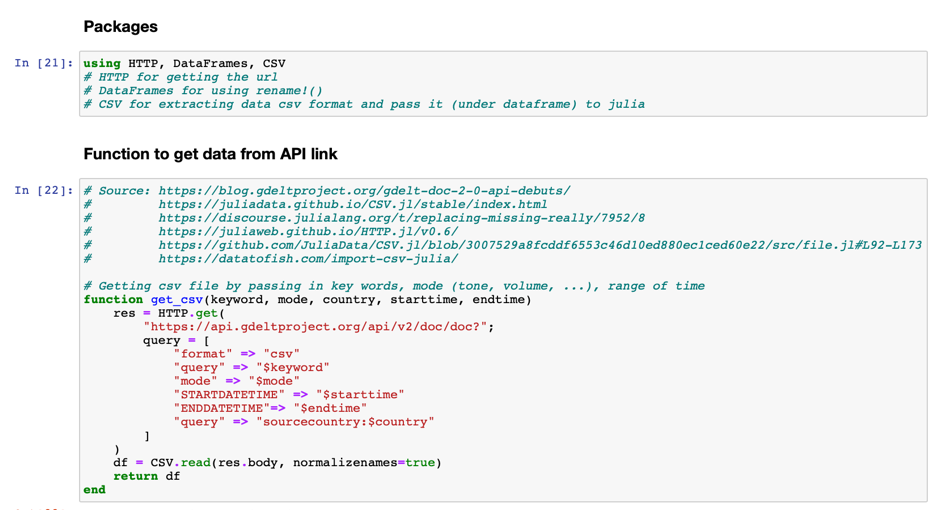
* + 1. **Techniques for webscraping**

|  |  |
| --- | --- |
| **R language** | |
| **Libraries** | **Comments** |
| Httr | For using API link |
| Magrittr | For using pipes |
| Tidyverse | For using rename() command |
| Glue | For using glue() command |
| **Julia language** | |
| HTTP | For getting the URL |
| DataFrames | For using rename!() command |
| CSV | For using CSV.read() command |

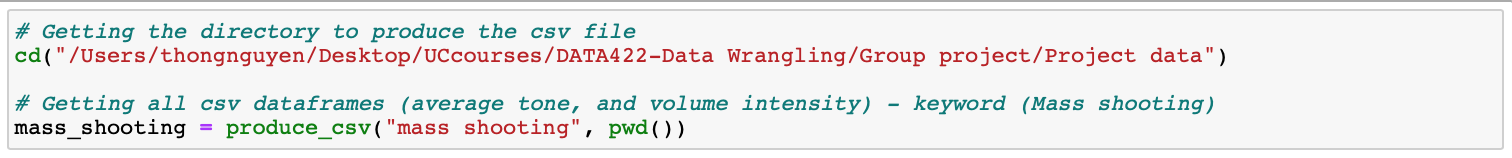
In both R and Julia, we built up two main functions for this section. In the first function, we used glue() to combine the url parts, and GET() to extract dataframe information from API link. The picture below would show how the first part was built.

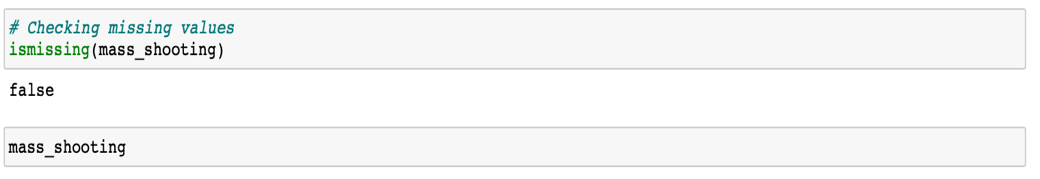
Following this, the first function was passed into the second function to obtain a list of dataframe under csv type using get\_csv() command. At this stage, get.csv() from tidyverse library was not used as this would produce some NA values as it could not read all the string types from the Gdelt project site. After that, write.csv() was used to save the csv file into our own local system (laptops) that we could do further modification on the dataset. Next, we added dir.create() command for storing all the dataframes under one folder which we could send to the group for each member working on their tasks. Before modifying steps, we checked again if any dataframes had missing values by is.na() and by viewing them. The picture below would show that process.



Secondly, the data of tone histograms for each year was extracted from Julia language. The process was done the same, except for addition about renaming the drop a column. In general, the details of the code would be mentioned in the code report.







* + 1. **Achievements and unachivements**

As to the achievements, from Gdelt Project website, we achieved the results which would need a lot of effort and a lot of time to complete. In particular, Gdelt helped us collect nearly all the newspaper around US writing about mass shootings, calculate the tone and show volume intensity. Furthermore, our ideas at the begining were to scrape the dataframe focusing on tone and volume intensity of the articles of each mass shooting. However, after researching about the Gdelt API document, we found another helpful information which was called ‘tonechart’ or tone histogram as mentioned above. This information could help us to see the distribution of tone about mass shooting articles around US each year.

As to the unachivements, the scraped data seems to have some urls that did not satisfy the conditions we have set up at the begining about the language and content. Specifically, the very small amount of them contained the other languages (not english) and contained irrelevant information about mass shooting. This is one of the minor mistakes of dataset from Gdelt API that it needs improving in the next version. Following that, this step would be improved in the further research on this project.

1. **References**

<https://www.nbcnews.com/storyline/las-vegas-shooting/las-vegas-shooting-deadliest-modern-u-s-history-n806486>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5296697/>

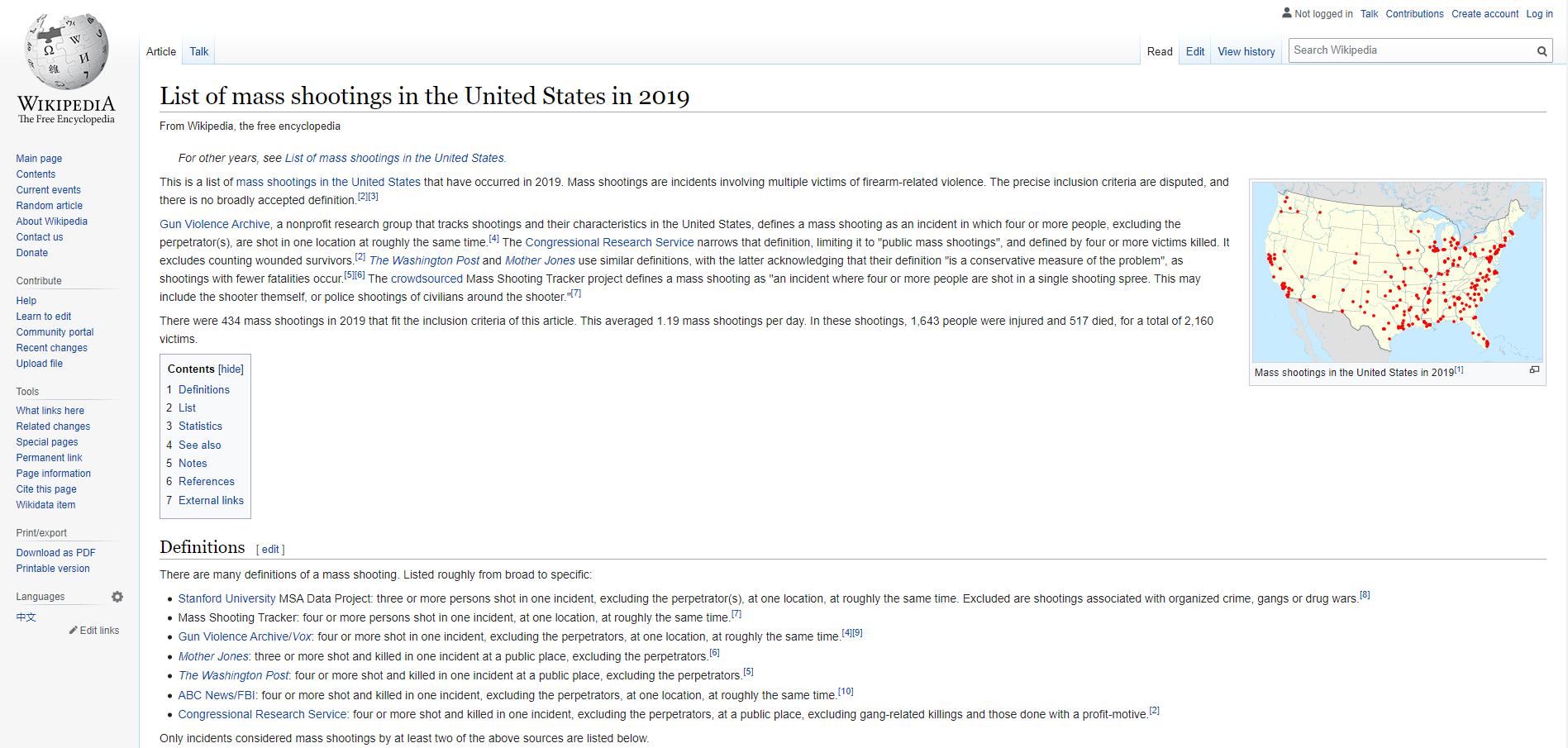
<https://www.gunviolencearchive.org/reports/mass-shooting?page=5&year=2018>

<https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/>

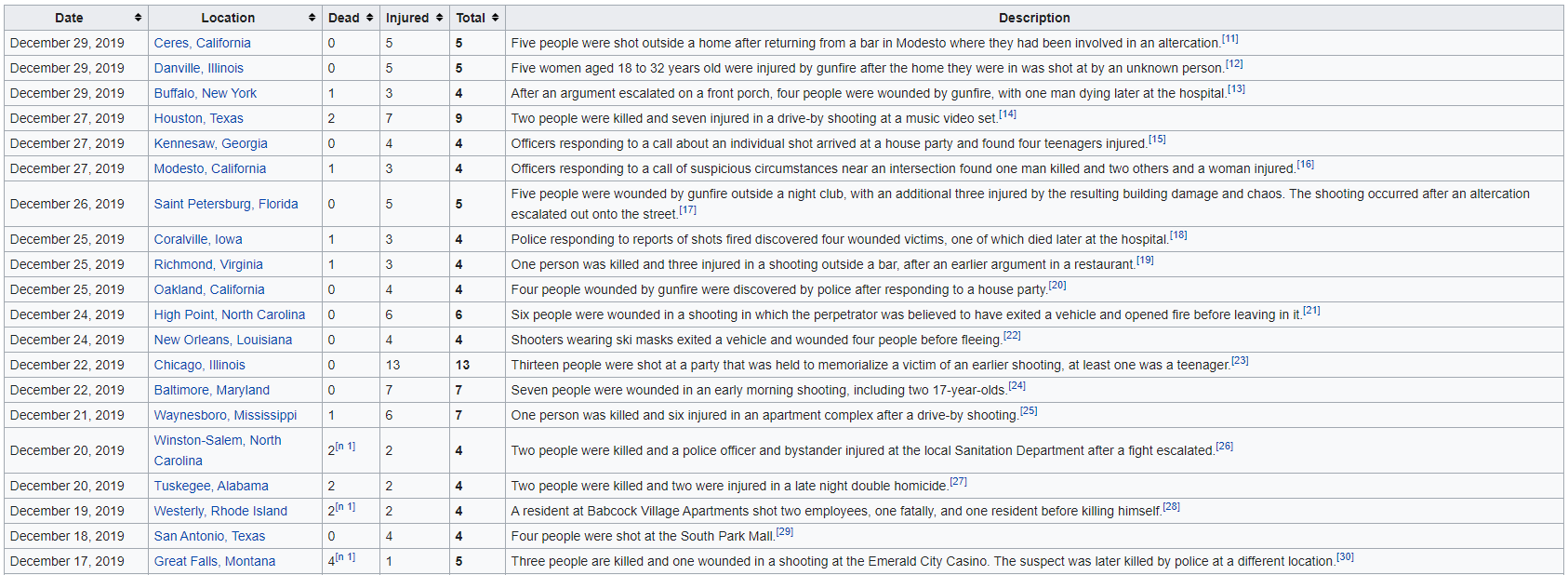
ANDY HUYNH – WIKI TABLE WEB-SCRAPING WITH R

We were able to find data about shootings that have happened in the US for 2018, 2019 and 2020. To build up the code from scraping the data from this source automatically, we have to try to write the code separately with small parts and make sure they were all working before combining them all under one function.

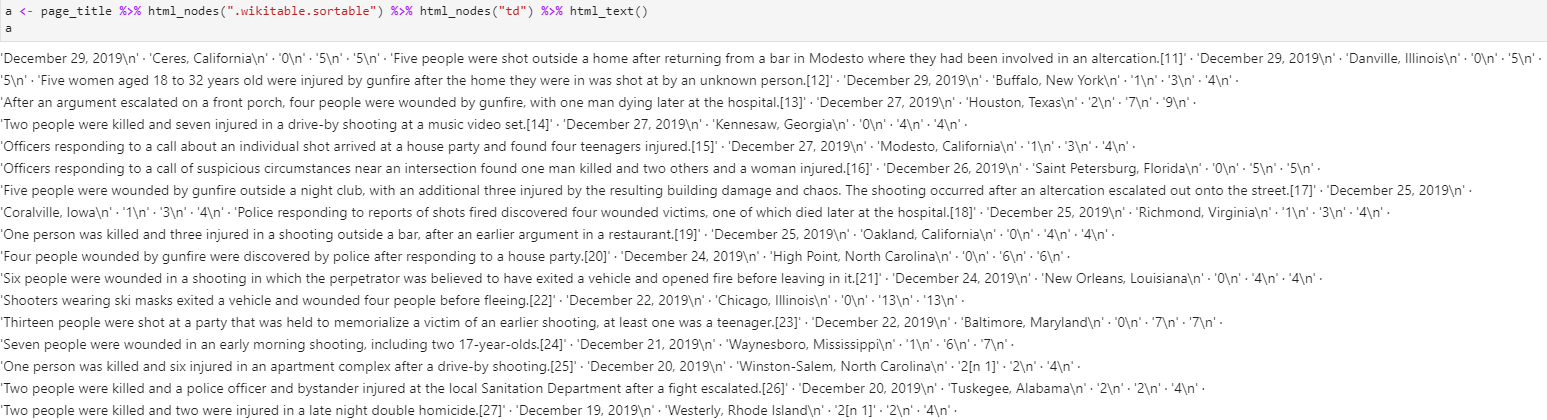
we firstly use Wikipedia page which contains the shooting in the US from the year 2019 at <https://en.wikipedia.org/wiki/List_of_mass_shootings_in_the_United_States_in_2019>



Including on this site, there is a data table which lists the shootings in the US as below:



So we decide to scrap the data from this table. We thought this would be an easy job, however very soon it turns out very challenging. After trying to find the node that contains the data in the table, we found out all kind of data under that table is included in the same node.



To have a better understanding of its structure. We use as.matrix() to view these data as a list.



From the result of this step, we studied that each row of the data under this node will contain a different kind of information. Such as the 1st row the Date, 2nd row is location, following rows are Death, Injured, Total and Description. And the 7th row will back to data of Date and so on. To solve this problem, we write a code with FOR loop and IF functions to split them into Different vectors named: Date, Location, Death, Injured, Total and Des.

After we got the vectors with the information that we wanted, we will create a new data-frame with columns are these vectors. As the data in the vectors are in the type of string, we have to convert Date to date-type before adding them to data-frame under Date column using mdy() under “lubridate” library. To fulfil columns of Death, Injured and Total in the data-frame, the vectors with this information must be converted into integer type by as.integer().



These steps of processing gave us new challenging. After successful combined all data we need into a data-frame, we found there are some NA values in Death, Injured and Total. Studied on the problem, we found that some of the number in the table on Wikipedia has some extension next to it, which is a sub-link to more details, and they are not a number. So they can not be converted to integers. To solve this problem, we have to remove these extensions before we can use these number as an integer type. Working on this, we notice that this extension starts with “[” and end with “]”. Therefore we use function rm\_between() under library(qdapRegex) to remove the square brackets and anything between them as well as extra “+”.



After successful remove all the unwanted information, we try to add them back to the data-frame as integers, and it returns without NA value. While we were checking the head(10) and tail(10) of the data-frame, there are some wrong information rows at the bottom of the data-frame. And this is caused by there is another data table under the node that we are working on.



To solve this problem. We used to use groupBy() to group all rows which contain for the year we are working on only and add them to the final output data-frame. Therefore the final data-frame will only have only rows with the correct year. Finally, we create the function() and a FOR loop to collect same kind data from 2018,2019 and 2020 and write them to CSV files.

We hoped that we would able to get the data of Shootings in the US from 2016 to 2020. However, we could only scrape the data from 2018 to 2020. This problem is not caused by there is no data from 2016 and 2017 on Wikipedia. The data for these two years is available in a different structure that our code can work on it automatically. Also, the information for 2016 and 2017 on this site is minimal, and would be misleading or not well present or what we managed to achieve.

# SAKTHI AND SYAIDA – WRANGLING AND VISUALISING DATA WITH R

# **Data Wrangling**

The target data model contains two different datasets: One is the clean shooting data combined with average tone and volume intensity and the second is the cleaned histogram tone data.

The R tidyverse library was used to wrangle the datasets into the target data model. Below are the different datasets and the wrangling process for each of them:

# **Shooting data from Wikipedia**

The shooting data was collected in three different csv files.

All three files contain 6 columns each with the 2018 shooting data file having 318 observations, the 2019 containing 438 observations while the 2020 file having 502 observations.

## Wrangling Involved

### Removal of the additional column

* All three files contained an additional column with the row numbers.
* As these columns were not used in any part of the data model, they were removed after data load.
* This was done by selecting the dataframe from the second column to last column which removes the first unused column.

### Date Type conversion

* The incident date column was of text data type in the files and loaded up as “Chr” datatype in R.
* So once the data was loaded, the date column in all three files were manually converted into date type.
* The conversion was done using the as.Date () method in R.
* This was done to work with the date column later.

### Combine the files into single dataframe

* All three shooting dataframes were combined into a single dataframe.
* This was done using the rbind method in R.

### Rename column

* The count column was renamed as “Victim Count” for easier readability.
* Using the colnames () method, the column name was changed using the index of the Count column provided.

### Filter shooting data for Mass shootings

* The dictionary doesn’t define a mass shooting precisely
* The Federal Bureau of Investigation’s (FBI) website states that any shooting with more than 3 deaths constitute as a mass shooting.
* For the purpose of this project, the FBI’s definition was followed and the shooting dataset was filtered for 4 or more deaths.
* This was done using the filter method on the “Death” column and the result was saved to the same shooting dataframe.

### Split location column into City and State

* The location column contains values for both City and State and is separated by a comma value.
* Using the separate () method, the two new columns called State and City were created from the location column that contained the values for the state and city respectively.
* The value of the separator was passed as “,” to the “sep” inside the separate method

### Remove empty spaces

* The State column created contained empty spaces for few values.
* They were removed using the trimws() method.
* This ensures the removal of whitespaces in both directions of the value.

### New column to capture the next day of the incident

* The new column was created to capture the next day of the incident
* This was done by adding a value of 1 to the “Date” column.

# **Average Tone data from GDELT**

The next dataset to wrangle is the Average Tone data from GDLET. This again was loaded from a CSV file. The CSV file contains 3 columns and 1013 observations.

## Wrangling Involved

### Removal of the additional column

* All three files contained an additional column with the row numbers.
* As these columns were not used in any part of the data model, they were removed after data load.
* This was done by selecting the dataframe from the second column to last column which removes the first unused column.

### Date Type conversion

* The incident date column was of text data type in the files and loaded up as “Chr” datatype in R.
* So once the data was loaded, the date column in all three files were manually converted into date type.
* The conversion was done using the as.Date () method in R.
* This was done to work with the date column later.

### Rename column

* The Value column was renamed as “Average Tone” for easier readability.
* Using the colnames () method, the column name was changed using the index of the Count column provided.

### Merging with the Shooting data Dataset

* The cleaned “Average tone” dataset needs to be merged with the “Shooting data” dataset to achieve the target dataset.
* This was done using the merge method: Both dataframes were passed into the method and an inner join was performed using the Date column from the Average tone dataset and the Nextday column in the Shooting data dataset.
* The required columns: *Date, Average Tone, City, State, Injured, Death, Victim Count and Description* were selected after the merge using the select method.

Visualizing the data

* From the merged dataset, a timeline of all mass shooting incidents was plotted with a bar showing the victim count in each shooting in the chart. This was done using the plotly package in R
* Another plot line showing the average tone for the day after the shooting was also plotted in the same chart.
* The victim count column was passed as the Y axis value and the Date column was passed as the X-axis value.
* Add trace method was used to the plot the individual plot lines in the same chart.
* A separate Y-axis was defined for the Tone plotline using the y-axis option inside the add\_trace method.
* The mass shooting description and the average tone was added to be displayed on hover over of the chart.

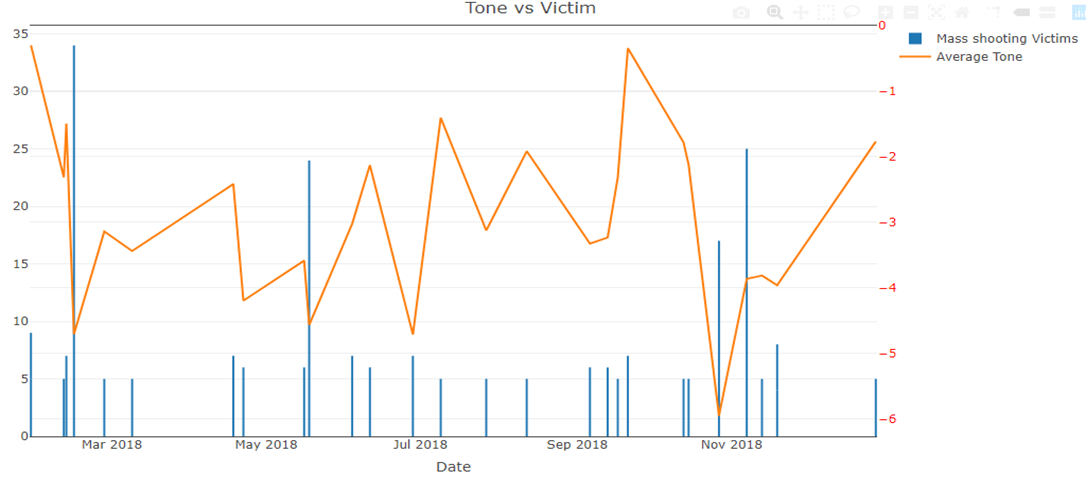


Fig 1. Average tone vs Victim count 2018

Inference

* From the graph above, Fig 1, it can be clearly seen that the victim count does not affect the average tone in the same way at all times.
* There are some shootings with a lower victim count that affect the tone more drastically than the ones with a higher victim count.

# **Volume Intensity data from GDELT**

The next dataset to wrangle is the Volume Intensity data from GDLET. This again was loaded from a CSV file which contained data for the years 2018-2020. The CSV file contains 14 columns and 626 observations.

## Wrangling Involved

### Removal of the additional column

* All three files contained an additional column with the row numbers.
* As these columns were not used in any part of the data model, they were removed after data load.
* This was done by selecting the dataframe from the second column to last column which removes the first unused column.

### Date Type conversion

* The incident date column was of text data type in the files and loaded up as “Chr” datatype in R.
* So once the data was loaded, the date column in all three files were manually converted into date type.
* The conversion was done using the as.Date () method in R.
* This was done to work with the date column later.

Combine the URL and Title columns to a list within a column (Wide to Long format)

* The columns containing URLs and Titles are identified using the contains method
* They are then nested into a list using the nest method with a key value.
* Then a mutate function is to create a new column from the output of the nest method.
* The result will provide two columns, URL and Title, that contains the list of all URLs and Titles respectively.
* The list will include the key and value combination.

### Merging with the Shooting data Dataset

* The cleaned “Volume intensity” dataset needs to be merged with the “Shooting data” dataset to achieve the target dataset.
* This was done using the merge method: Both dataframes were passed into the method and an inner join was performed using the Date column from the volume intensity dataset and the Nextday column in the Shooting data dataset.
* The required columns: *Date, Volume Intensity, City, State, Injured, Death, Victim Count, Description, URL and Title* were selected after the merge using the select method.

Visualizing the data

* From the merged dataset, a timeline of all mass shooting incidents was plotted with a bar showing the victim count in each shooting in the chart. This was done using the plotly package in R
* Another plot line showing the volume intensity for the day after the shooting was also plotted in the same chart.
* The victim count column was passed as the Y axis value and the Date column was passed as the X-axis value.
* Add trace method was used to the plot the individual plot lines in the same chart.
* A separate Y-axis was defined for the Tone plotline using the y-axis option inside the add\_trace method.
* The mass shooting description and the volume intensity was added to be displayed on hover over of the chart.

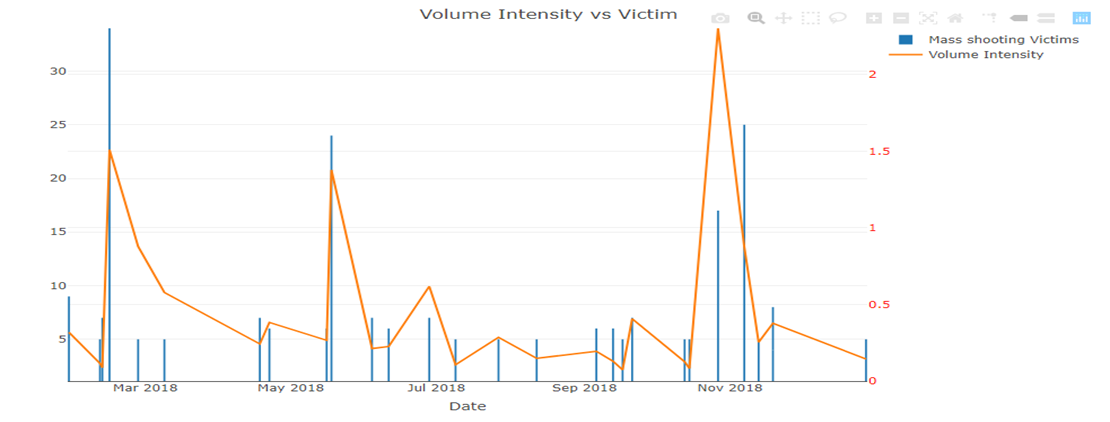


Fig 2. Volume Intensity vs Victim count 2018

Inference

* From the graph above, Fig 2, it can be clearly seen that the victim count does not affect the volume intensity in the same way at all times.
* There are some shootings with a lower victim count that affect the intensity more drastically than the ones with a higher victim count.

**Merging Shooting data, Volume Intensity and Average tone to obtain the final dataset**

* The two merged datasets were combined to obtain the final clean dataset
* The datasets were combined using the date, city and state columns.
* The required columns : Date, Average Tone, Volume Intensity, City, State, Injured, Death, Victim Count, Description, URL and title were selected and the clean dataframe was generated.

**HISTOGRAM TONE DATA**

This dataset consists of details regarding the title and URL of the top 10 articles related to mass shootings for the year 2018 to 2020 in United States. All these datasets contain 23 columns with the year 2018 dataset containing 53 rows, 2019 containing 49 rows and the year 2020 with 44 rows of observations.

## Wrangling Involved:

### Date type column conversion

The date type was originally in a form of character type, therefore it was changed to a date type.

### Remove X column in all the dataframes

Since the X column seems to be not very useful for the project as it only consists of numbers representing the row numbers, therefore the column was removed by selecting only the columns after the first column of the dataset.

### Renaming the columns

The column names were then changed to a more meaningful name for easier readability.

### Gather histogram data for each of the years 2018, 2019 and 2020.

The nest () function was used to create two list-columns of data frames whereby the first one includes only information with URL and the other one including the titles of the articles.

The mutate () function was then used to create two new columns and to separately include the URLs and the titles.

Visualizing the data:

The below histograms represent the count of articles for each tone values presented by the articles for the year 2018, 2019 and 2020.

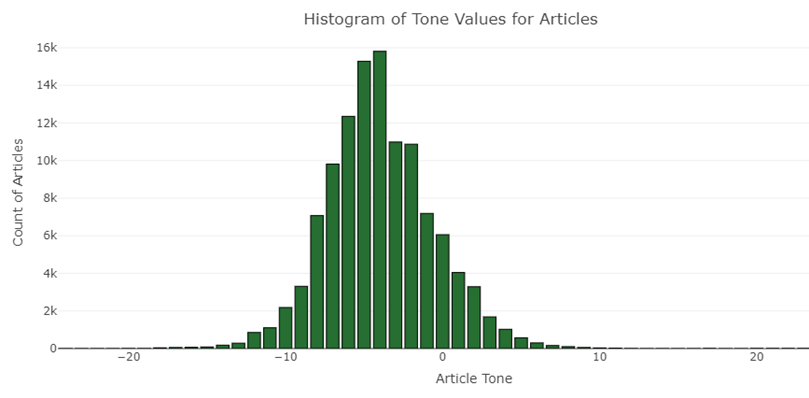


Fig 3 Histogram of Tone: 2018

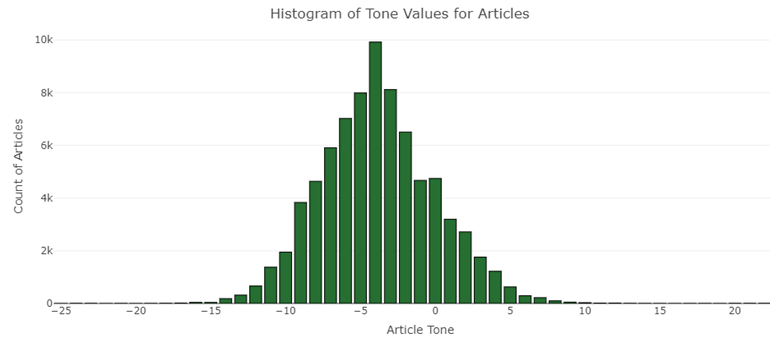


Fig 4 Histogram of Tone: 2018

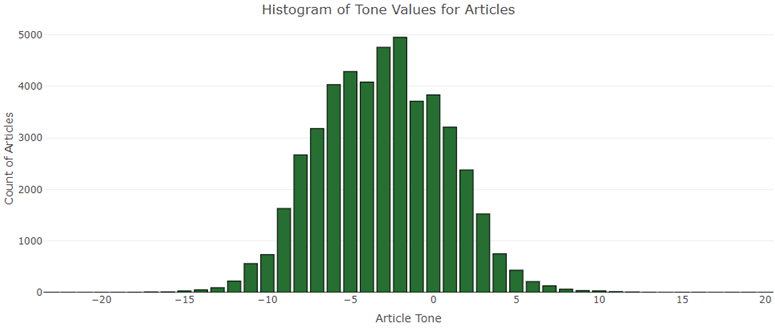


Fig 5 Histogram of Tone: 2018

From here we could observe that as expected the number of articles that represent a certain tone over the three years fall below showing that mostly they are negative.

Function creation for easier access in future

Functions were created allow easier access for the final datasets so that anyone coming in new can access the data quickly. All functions created accepts the first input parameter “Type” as Date, State or City and then a second parameter for the value. Below were the functions created:

* Function to return Average Tone by Victims Dataframe based on Date/State/City
* Function to return Volume Intensity by Victims Dataframe based on Date/State/City
* Function to return Histogram Tone Data based on based on Date/State/City
* Function to return the final dataset based on Date/State/City

Challenges faced:

Among the challenges faced, the major one was when having to apply the nest method which was used to bring all the title and URL columns into a list. The solution was difficult to obtain and various methods were tried before coming to use the nested method. Since the original dataset had several columns representing URLs and titles of the article, this was a really important issue to solve to proceed further with the project. The dataset was then transformed from a wide format into a long type. The other major challenge was in terms of having to visualize a single plot that contains two y- axis in the same graph since each y axis represented different variables. This was achieved using the y-axis method in the add\_trace function in plotly.

Future Extension:

* One of the major feedbacks from the project presentation was that the Average tone value is only picked up for the next day after the incident and it does provide a complete picture. This needs to addressed in future expansions, by including a range of days before and after the shooting incident.
* The project currently includes only the shooting data for United States. In future, this can be extended to other countries in the world as well.
* The shooting data can be picked up for other historical years as well. This will provide more interesting patterns
* Based on the current analysis, the project can be extended further to analyze why exactly the tone and volume intensity vary differently. This could be due to various reasons and analyzing them will reveal more interesting patterns.